Concordances et concordanciers : de l'art du bon KWAC

Bénédicte Pincemin

in Corpus en Lettres et Sciences sociales : des documents numériques à l'interprétation, *Actes du XVIIe Colloque d'Albi Langages et Signification, Albi, 10-14 juillet 2006, Carine Duteil-Mougel & Baptiste Foulquié (éds), ISBN 2-907955-12-18, pp. 33-42, et Texto! [en ligne http://www.revue-texto.net/ ISSN 1773-0120], juin 2006, vol. XI, n°2.*

CONCORDANCES ET CONCORDANCIERS DE L'ART DU BON KWAC

Bénédicte PINCEMIN
CNRS / Université de Paris 13, LLI

SOMMAIRE

Synthèse : Les concordances sont un mode de présentation d'extraits de texte, contenant tous le même mot ou le même motif linguistique. C'est une méthodologie d'analyse textuelle séculaire, l'exemple même du "travail de bénédictin" avant le recours possible aux ordinateurs. Les concordanciers sont des outils informatiques produisant les concordances souhaitées à partir d'un corpus numérique. Le sujet qui nous intéresse est de comprendre cette remarquable pérennité de la concordance, et de repérer et analyser les transformations discrètes mais décisives apportées par leur calcul automatisé. Ce parcours conduit également à s'interroger sur une exploitation plus systématique et pertinente des possibilités ouvertes par l'ordinateur : quelle généralisation opératoire peut-on définir de la méthode des concordances ?

Classiquement, dans les logiciels d'analyse textuelle, un calcul de concordance se définit par la détermination de trois paramètres : la donnée d'un pivot, à savoir le mot ou motif linguistique dont on veut étudier les occurrences en contexte ; la taille du contexte à visualiser ; et un critère de tri (éventuellement multiple) fixant l'ordre de présentation des contextes. Notre généralisation est spécifiante : plutôt que de multiplier les paramétrages et réglages, nous proposons de focaliser les ajustements sur ce qui fait la force des concordances calculées. Nous retenons donc pour une "bonne" concordance deux caractéristiques essentielles : les contextes présentés sur une ligne et superposés, de sorte à créer des effets visuels d'alignement vertical, d'une part ; et le tri des lignes de contexte, sur un ou plusieurs éléments, d'autre part.

Chemin faisant, nous écartons donc clairement d'autres relevés d'occurrences quelquefois appelés concordances. En nous inspirant très librement de désignations de types d'index en sciences de l'information, nous distinguons le KWOC (keyword out of context), liste d'attestations ; le KWAC (keyword and context), concordance telle que nous l'entendons, avec son dispositif de regroupement visuel par superposition, alignement vertical et tri ; le KWIC (keyword in context), relevé de contextes avec un choix plus libre de la taille de ceux-ci ; le KWUT (keyword up to text), où les occurrences sont repérées au fil du texte. Ces quatre modes de relevé d'occurrences sont complémentaires et gagnent à être cultivés et utilisés pour leurs spécificités.

La concordance "papier" traditionnelle et les sorties d'un concordancier diffèrent davantage que par leur mode de production ; mieux, chacune tire le meilleur parti des spécificités de leur processus de construction :travail de synthèse et guide interprétatif pour la première, régularité, polyvalence et dynamique pour la seconde. Cependant, par des voies différentes, concordances manuelles et concordances calculées servent le même principe herméneutique fondamental : la mise en évidence des parallélismes et des contrastes dans les contextes de l'item étudié.

Dans cette optique, pour la concordance (KWAC), le réglage de la taille des contextes devient secondaire (voire nuisible, car dénaturant potentiellement la concordance ; c'est plutôt une caractéristique du KWIC). A l'inverse, le mode de définition du pivot gagne à être affiné, par l'introduction d'une décomposition en zones, qui permettent de démultiplier et d'assouplir les dispositifs d'alignement et de rapprochement visuels (Pincemin et al. 2006).

Mots-clés : concordances, statistique textuelle, herméneutique.

1. Comprendre le succès des concordances

Il est frappant d'observer la popularité toujours actuelle des concordances, telle que manifestée par l'abondance des concordanciers[1] et l'omniprésence de la fonction de calcul de concordances dans les logiciels d'analyse de texte. Pourtant, la procédure informatique est simple, relativement à d'autres techniques d'analyse textuelle : il s'agit d'une réorganisation du matériau textuel par une indexation machinale (relevé systématique des mots[2]), un tri formel (alphabétique), et une présentation astucieuse. L'état de l'art des outils d'analyse textuelle révèle et démontre toute une gamme de procédures plus élaborées, tout particulièrement dans le domaine des statistiques textuelles (lexicométrie puis textométrie)[3]. Mais ces techniques sont encore peu diffusées, restent surtout l'apanage de logiciels de recherche français (tels que Weblex, Hyperbase, Lexico). Les concordances restent de fait l'outil de référence hors de la communauté des statistiques textuelles, comme au plan international, pour les outils de consultation et d'analyse de corpus numériques.

Ce succès témoigne à la fois de la complémentarité de cette technique relativement aux développements innovants de la statistiques textuelle, et, dans l'absolu, du bien-fondé herméneutique, de la pertinence, de l'efficacité des concordances pour l'étude des textes. D'où l'intérêt d'examiner de plus près leur conception et leur fonctionnement : sur quels principes les concordances sont-elles fondées ? Quelles sont les éventuelles variantes de réalisation ? Y a-t-il des versions plus intéressantes que d'autres, pour quelles raisons ou/et sur quels plans ?

2. Des trois paramètres fondamentaux des concordanciers à une définition des concordances

L'expérience du recours à différents logiciels permet d'abord de repérer une définition concrète de la concordance, à travers les paramétrages prévus. En pratique, une concordance se calcule à partir de trois éléments :

(i) un pivot, typiquement un mot (mais cela peut être généralisé à toutes sortes d'items dont les occurrences sont repérables par le logiciel considéré : lemme, expression, etc.), aux contextes duquel on s'intéresse ;

(ii) une taille de contexte (éventuellement détaillée en contexte gauche et contexte droit, qui respectivement précède et suit le pivot), typiquement la longueur d'une ligne pour le mode d'affichage utilisé ;

(iii) un ordre de présentation des contextes sélectionnés, typiquement ordre de présence dans le corpus, ou tri alphabétique sur le mot qui précède le pivot (tri gauche) ou sur celui qui le suit (tri droit).

Cette première généralisation rend ensuite sensible aux implémentations qui restent en-deçà de ces possibilités, ou à l'inverse qui étendent la portée d'un paramètre. Dans le premier cas (implémentation pauvre), l'incidence de l'occultation d'un paramètre n'est pas forcément négative, si elle simplifie l'application informatique en la centrant d'emblée sur des valeurs de paramètre suffisantes et efficaces pour les domaines d'usage prévus. Par exemple, pour l'étude linguistique du lexique telle que la pratique notre laboratoire, selon une approche de type sémantique distributionnelle (le sens d'un mot est caractérisé par ses contextes d'emploi, notamment par ses dépendances syntaxiques), l'absence du second paramètre (si le logiciel affiche toujours des contextes d'une ligne, de longueur fixe, optimale) s'avère souvent moins pénalisante que l'impossibilité de trier les contextes, car les alternatives permises par ce troisième paramètre répondent à une plus grande diversité de questionnements.

De fait, la définition par les paramètres laisse implicite une caractéristique essentielle des

---

[1] Par exemple : AntConc, KWICFinder, MonoConc, TACT…

[2] La linguistique montre toute la complexité de la notion de "mot" ; on s'en tient ici aux définitions opératoires, intuitives et simplifiées, qui sont utilisées en statistique textuelle, en termes de chaînes de caractères non délimiteurs maximales (Lebart & Salem 1994).

[3] Sans même entrer ici dans le domaine, très développé et actif, des analyses faisant appel à des modélisations linguistiques ou/et sémantiques.

concordances : l'alignement vertical, sur une colonne (généralement centrée), des occurrences du pivot. Cet alignement en colonne, associé au tri des lignes de contexte permettant de rapprocher les contextes analogues, souligne visuellement, par superposition et répétitions, les convergences et les divergences de formulation. Cette présentation s'avère un outil heuristique extrêmement efficace pour la lecture des contextes proches et l'observation globale des constructions dans lesquelles s'insère le mot étudié.

Ainsi, alors que l'usage désigne par KWIC (KeyWord In Context) la technique des concordanciers, nous préférerions parler ici de KWAC (KeyWord And Context), pour souligner que les relevés de contextes sont organisés autour du mot étudié et à partir de lui. La concordance est concordance parce qu'elle est visuellement ancrée sur le pivot qui structure le parcours de lecture. La disposition invite à faire du mot-clé (*KeyWord*) le point d'entrée ou du moins un repère central constant, et (*And*) son contexte (*Context*) immédiat est disposé de part et d'autre. Ce faisant, la désignation de KWIC reste disponible pour distinguer une forme complémentaire de relevé d'occurrences, pour laquelle le mot (*KeyWord*) reste davantage inséré (*In*) dans son contexte (*Context*), la présentation lui donnant un rôle moins dominant (cf. § 3).

La pratique des concordanciers rend enfin sensible à l'importance interprétative de l'indication, pour chaque contexte extrait, d'une référence intelligible permettant de le situer au sein du corpus. En effet, l'interprétation se nourrit d'informations locales (les mots dans le voisinage étudié) mais aussi globales : par exemple, de quel texte est tiré ce contexte ? à quel moment a-t-il été écrit, de quelle source provient-il ? Autant d'indications sans lesquelles l'interprétation, artificiellement privée d'une grande part du contexte, est considérablement appauvrie. Les concordanciers peuvent maintenant recourir utilement à l'hypertexte pour donner directement accès au point du texte dont provient le contexte. Ceci ne périme cependant pas une mention mnémonique, choisie pour apporter les informations intertextuelles jugées utiles, et que l'on a sous le regard en même temps que l'on balaye les occurrences en contexte. Ces références de localisation se contextualisent en outre mutuellement : on y lit le passage d'une partie du corpus à une autre, la richesse ou la pénurie d'attestations selon les localisations.

Ces observations permettent de formuler une définition synthétique du (bon) KWAC : Un corpus étant fixé, une concordance (ici KWAC) est la liste de toutes les occurrences d'un pivot, alignées verticalement en colonne, entourées de part et d'autre par leur contexte, accompagnées d'une référence indiquant de façon pertinente leur positionnement dans le corpus, et triées selon un critère pertinent pour l'analyse.

3. KWOC, KWAC, KWIC et KWUT : Une typologie des relevés d'occurrences

Plus généralement, une certaine confusion règne entre différentes formes de relevés lexicaux au sein d'un texte numérique. En nous inspirant très librement d'une formulation mnémonique ayant cours dans les sciences de l'information (cf. § 5), nous proposons de différencier les KWOC, les KWAC, les KWIC et les KWUT.

Les KWOC sont les KeyWord Out of Context. Il s'agit de la liste des différentes attestations correspondant à la requête soumise par l'utilisateur, qui fait office de filtre sur le vocabulaire complet. Comme il n'y a pas de contexte (sinon celui, interne, de l'occurrence elle-même, qui peut être une expression composée), les KWOC ne se laissent guère confondre avec les concordances et autres relevés de contexte, même s'il y a une certaine parenté théorique et technique entre ces procédures. La force du KWOC vient de son caractère synthétique. L'effacement du contexte permet une réduction supplémentaire à celle d'une concordance à taille de contexte nulle : c'est le regroupement des différentes occurrences ayant la même forme. Pour un même choix de pivot, un KWOC est donc d'une longueur bien moindre que celle d'une concordance, en évitant certaines redondances si l'attention se porte moins sur le contexte que sur le pivot. En chiffrant les répétitions condensées sous une même forme et en offrant des tris tant sur les formes (alphabétique, canonique) que sur les fréquences (tri dit hiérarchique), les relevés KWOC sont très efficaces pour jauger le relief quantitatif de différentes formulations, pics attracteurs comme lacunes surprenantes et significatives (l'équipe Hubert de Phalèse a même trouvé un nom à ces dernières (fréquences nulles) : les *nullax*, de la même manière qu'on a le cas particulier remarquable des hapax, occurrences de fréquence 1). Enfin, remarquons que ce "hors contexte" n'est en rien décontextualisé, le contexte immédiat ne s'éclipse que pour mieux manifester le contexte textuel ou intertextuel, grâce aux références de localisation, qui résument le profil de répartition des occurrences. Malgré son nom, le relevé KWOC est pleinement lié au corpus, dont il pointe efficacement des caractéristiques expressives profondes et pas toujours perceptibles.

Les KWAC sont les KeyWord And Context. Les occurrences du pivot étudié sont présentées avec leur contexte et alignées verticalement pour former une colonne, de telle sorte qu'une double lecture est possible, verticale (liste des formes du pivot : KeyWord) et (And) horizontale (contextes : Context). Cette superposition est essentielle car elle met en évidence les répétitions plus ou moins massives et les variantes, tout particulièrement lorsqu'elle est associée à un tri alphabétique des contextes. Elle donne appui à un parcours de lecture des voisinages d'un mot complètement nouveau et très riche, mettant très efficacement en lumière des régularités d'usage pas toujours perceptibles dans une lecture linéaire classique. C'est aux relevés présentant une telle heuristique visuelle de lecture, par "empilement", que nous réservons l'appellation de concordance. Comme nous l'avons vu, selon cette perspective, la définition pratique des concordances par les trois paramètres -pivot / taille du contexte / ordre de présentation- doit être révisée. S'ajoute une caractéristique définitoire : l'alignement vertical des occurrences du pivot. La définition en est alors corrigée : l'effet d'empilement ne se réalise pleinement que si les contextes se présentent chacun sur une ligne, la taille du contexte n'a alors plus vraiment lieu de rester un paramètre.

Les KWIC sont les KeyWord In Context[4]. Il s'agit d'un relevé des contextes d'un mot, en privilégiant souvent des contextes assez larges, de l'ordre de plusieurs lignes ou phrases, ou d'un paragraphe par exemple. Le mot est mis en valeur afin de donner appui à cette résonance interprétative entre le mot (KeyWord) et son contexte (Context). Une telle taille de contexte, élargie et assouplie, annule en grande partie les effets de lecture permis par un alignement vertical sur le pivot et une superposition des contextes, qui reste l'apanage du KWAC. On retrouve ainsi les deux premiers paramètres des concordances (choix du pivot et de la taille des extraits), mais le troisième n'est plus toujours pertinent. La force du KWIC, c'est l'optimisation potentielle des contextes locaux d'observation : toute liberté est donnée sur la taille ou la nature des contextes, et ceux-ci peuvent être restitués selon leur présentation usuelle (en préservant par exemple ainsi l'information visuelle de l'occurrence en début / fin de paragraphe). Lorsqu'un même contexte contient plusieurs occurrences, il n'y a évidemment pas lieu de le répéter à l'identique. Les concentrations locales d'occurrences (répétition du pivot) sont donc particulièrement bien rendues (visualisation de la disposition dans le contexte local), sans alourdir inutilement le relevé. Un "bon" KWIC a une très forte valeur interprétative : il s'agit d'un recueil de contextes délimités de façon pertinente, de véritables unités sémantiques pour l'interprétation des attestations, bénéficiant d'une certaine autonomie. Chaque contexte peut devenir comme un petit texte : ce n'est pas sans rappeler la pratique des recueils de citations, ou des morceaux choisis. Sans renoncer au contexte du corpus (normalement motivé, par construction), l'interprétation peut s'enrichir en jouant de focalisations à géométrie variable, à la fois souples et contrôlées (tout découpage n'est pas pertinent).

C'est entre les KWIC et les KWAC que les confusions sont les plus répandues : le KWIC pourrait n'être vu que comme un KWAC au contexte plus long ; et, dans les concordanciers usuels, le réglage de la taille de contexte pourrait être un mode de basculement insensible du KWAC au KWIC. En fait, KWIC et KWAC sont deux spécialisations complémentaires du concordancier, donnant chacune toute sa mesure à l'un des paramètres, sans que l'un bride ou affaiblisse l'autre : le KWIC libère l'éventail des tailles et types de contextes, et le KWAC tire le meilleur parti des effets visuels, heuristiques et herméneutiques, offerts par les tris. Les contextes sur une ligne et superposés du KWAC se prêtent bien à une organisation selon une progression continue, et les contextes plus textuels du KWIC se structurent assez naturellement en classes, par regroupements, rapprochements et contrastes.

A la différence du KWAC qui est par nature centré sur le voisinage immédiat du pivot, le KWIC

[4] KWIC est une formule plus courante que les deux autres (KWOC et KWAC -KWUT est un néologisme de notre crû), et est souvent utilisée pour désigner les concordances. Nous nous écartons donc de l'usage, pour tirer bénéfice de la typologie soulignée par la formulation mnémonique quadruple KWIC KWAC KWOC KWUT.

permet d'explorer un contexte focalisé tout en étant délimité plus souplement. Le KWIC peut ainsi être une réponse bien adaptée à certains questionnements sémantiques par exemple, lorsque les incidences sémantiques recherchées sont diffuses et peu cadrées, alors que le KWAC pourra être plus efficace pour l'étude de relations ancrées dans le lexique ou la syntaxe, à portée plus limitée et aux formes plus régulières, mieux saisissables par des tris. Un cas intéressant est celui de la technique de concordance enrichie mise au point par Evelyne Bourion (2001). Dans une perspective d'analyse sémantique, les mots statistiquement significativement associés au pivot (les corrélats) sont mis en valeur typographiquement ; les lignes de concordance sont triées en fonction de la densité et de la force de ces associations sémantiques potentielles calculées. Or ces corrélats ne créent que peu ou pas d'effets d'alignement visuels, leur relation, sémantique, est souple en terme de construction syntaxique comme de distance au pivot. Ces relevés sont donc un cas limite de KWAC : l'effet de superposition est affaibli (pas d'empilement en colonne des corrélats), cependant présent (si le contexte est sur une ligne et grâce à la mise en valeur typographique qui attire le regard et facilite l'établissement de correspondances).

Les KWUT sont les KeyWord Up to Text. Le mot étudié est mis en valeur au sein de son contexte, au fil du texte. Autrement dit, le texte est affiché dans son entier, dans son déroulement intégral, et les occurrences sont repérées dans ce contexte global. Le KWUT donne accès à l'organisation des occurrences à l'échelle textuelle, à leur disposition dans la linéarité textuelle. Il peut par exemple y avoir des effets de regroupement dans une zone du texte, positionnée à tel endroit (début, fin du texte,...) ou à l'inverse d'évitement de telle ou telle partie. Comme on revient au texte dans son intégralité avec sa mise en forme, a minima le marquage de ses grandes subdivisions, le KWUT renseigne aussi sur la saillance éventuelle de telle ou telle occurrence du fait de sa position, notamment aux frontières : par exemple mot dans la phrase qui débute un chapitre, ou "mot de la fin". Pour être exploitable, la présentation KWUT devrait être couplée à un dispositif de repérage rapide des occurrences, de sorte à ne pas être contraint de parcourir linéairement tout le texte pour visualiser toutes les occurrences. Généralement on dispose d'une fonctionnalité de saut d'une occurrence à l'occurrence suivante (ou précédente). Une autre possibilité, complémentaire car plus textuelle, mais encore peu répandue, est la présentation conjointe (et hypertextuellement liée) d'une vue d'ensemble de la distribution des occurrences du pivot dans le texte. La "carte des paragraphes" du logiciel Lexico 3, ou l'histogramme marginal proposé par Pincemin[5], sont des exemples de représentations textuelles visuelles et synthétiques de ce type.

A ce parcours du KWOC au KWUT, s'associe une extension progressive des contextes : lexie et syntagme pour le KWOC, syntagme à période avec le KWAC, période ou paragraphe avec le KWIC, texte voire intertexte dans le KWUT. L'intertexte ne semble pas motiver une forme de relevé en contexte supplémentaire, mais il s'introduit de façon plus ou moins saillante dans les autres relevés. Un KWOC peut détailler la répartition des formes ou/et des fréquences dans les différentes parties du corpus (textes, catégories), et donner ainsi une vision globale des attestations par rapport à l'articulation intertextuelle du corpus. Si le corpus présente une organisation orientée, les KWAC sont quant à eux bien appropriés pour rendre compte synthétiquement des effets de succession et d'évolution, grâce au tri des contextes selon l'ordre du corpus. Les contextes KWIC se prêtent particulièrement bien à une structuration par regroupements : les divisions intertextuelles peuvent éclairer l'analyse des relevés KWIC, et réciproquement les contextes d'attestation KWIC peuvent caractériser les articulations du corpus quant à l'usage de telle ou telle forme.

4. Illustration : propositions pour les concordanciers sur corpus multilingues parallèles alignés

Considérons le cas où l'on dispose d'un corpus numérique formé de différentes versions d'un même texte, traduit en plusieurs langues. Les correspondances d'une langue à l'autre sont accessibles entre passages plus ou moins fins : selon la nature du texte et la procédure d'alignement cela peut être de l'ordre du paragraphe, de la phrase ou du mot.

[5] Voir par exemple Pincemin B. (2001). « Résoudre la surcharge informationnelle sans décontextualiser », in Stéphane Chaudiron et Christian Fluhr (éds), 3ème colloque du chapitre français de l'ISKO « Filtrage et résumé automatique de l'information sur les réseaux », Université Paris X, 5-6 juillet 2001, pp. 149-158.

Le KWOC peut être utile par exemple si l'on dispose d'une forme d'alignement au niveau des mots, pour lister synthétiquement les correspondances lexicales d'un mot dans une autre langue dans le contexte du corpus étudié.

Le KWAC est très efficace pour étudier les usages d'un mot à l'intérieur d'une langue, notamment les constructions grammaticales dans lesquelles il s'inscrit, et ses associations sémantiques proches (comme les choix des qualificatifs).

Pour l'étude multilingue et simultanée, parallèle, des contextes d'un mot, un KWIC semble plus pertinent qu'un KWAC, car l'effet de superposition et de tri est très affaibli et souvent non pertinent lorsque les langues sont mêlées ou intercalées. Une présentation efficace serait donc un KWIC sur plusieurs colonnes (une par langue) relativement étroites, de sorte à donner de l'épaisseur visuelle aux contextes, et faciliter ainsi l'observation des correspondances et des décalages de positionnement, eux-même révélateurs des variantes de formulation.

Autrement dit, un concordancier multilingue n'est pas (ne devrait pas être) la simple application d'un concordancier usuel à un corpus multilingue. Mais aussi, une concordance sur corpus aligné ne gagne rien à être un alignement de concordances. Même en contexte multilingue, la concordance se centre et se calcule sur une langue. Cette affirmation n'exclut pas bien sûr la possibilité de consulter facilement, depuis la concordance, tel contexte aligné dans telle autre langue ; elle n'exclut pas non plus la possibilité de calculer, en se basant sur l'analyse d'une concordance dans une langue, une autre concordance dans une autre langue, avec des facilités pour les visualiser conjointement et naviguer de l'une à l'autre.

5. Retour épistémologique et terminologique sur les KWIC

Notre proposition de typologie des relevés d'attestations en KWOC, KWAC, KWIC et KWUT a l'avantage d'être mnémonique, tout en contrastant bien différents modes de présentation complémentaires. Mais elle a l'inconvénient d'aller à contre-courant de l'usage particulièrement bien établi et largement diffusé pour les KWIC, et des notions apparentées de KWOC et de KWAC, classiques bien connus des experts en gestion de l'information. Il nous semble nécessaire de rendre compte ici de ce contexte épistémologique.

Le concept de KWIC, et sa désignation, ont été définis par Hans Peter Luhn, à la fin des années 50. Il s'agissait de construire un index des titres de publications scientifiques, selon une procédure simple automatisable (afin de disposer d'un index très complet actualisé très fréquemment). Les mots-clés sont alors les mots lexicaux des titres (on écarte simplement les mots grammaticaux et éventuellement des mots généraux jugés ici peu significatifs). Ces mots sont alignés en colonne au centre de la page, entourés de part et d'autre par leur contexte dans chaque titre. Ces relevés sont triés selon les mots-clés centrés, de façon à former l'index : les différents mots-clés apparaissent alors dans l'ordre alphabétique, avec pour chacun le relevé des différents titres dans lequel il apparaît. Voici une illustration de ce procédé appliqué à deux intitulés de colloque :

```
SdN06                          Semaine du    Document    numérique
Albi06                           Colloque    Documents   numériques et interprétation
Albi06   Colloque Documents numériques et interprétation
SdN06                 Semaine du Document    numérique
Albi06                 Colloque Documents   numériques   et interprétation
```

Fig. 1 : Exemple de KWIC (au sens traditionnel)

Le procédé consiste donc à faire tourner astucieusement les chaînes de caractères des titres ("rotated strings"). Le KWOC et le KWAC sont alors des variantes à partir de ce principe[6].

[6] D'après les documents que j'ai pu consulter (en particulier
Hjørland B. (2006). KWIC / KWAC / KWOC.
http://www.db.dk/bh/lifeboat_ko/SPECIFIC%20SYSTEMS/kwic_kwac_kwoc.htm (28/01/2006)
qui renvoie lui-même à
Buckland M. (2003). Organization of Information in Collections : Verbal Access.
http://www.sims.berkeley.edu/courses/is245/s03/verbal.html
et

Le KWAC (KeyWord Alongside Context, ou KeyWord And Context[7]) a pu être imaginé pour optimiser le volume. En effet, l'index KWIC sur les titres (contextes courts cités dans leur entier) n'occupe en moyenne qu'une moitié de chaque ligne[8]. Pour éviter ces blancs liés au centrage du mot-clé, le mot-clé peut être placé en tête de ligne, suivi de son contexte ultérieur. Puis, après un séparateur conventionnel (éventuellement tout simplement le point), le contexte antérieur est donné. Cette tactique traduit très directement cette idée de « chaînes tournantes » (rotated strings). Outre le gain de place, l'association entre le mot-clé et sa référence de localisation est facilitée. Voici une illustration du KWAC sur les mêmes données que le KWIC précédent :

```
SdN06    Document numérique. Semaine du
Albi06   Documents numériques et interprétation. Colloque
Albi06   interprétation . Colloque Documents numériques et
SdN06    numérique . Semaine du Document
Albi06   numériques et interprétation. Colloque Documents
```

Fig. 2 : Exemple de KWAC (au sens traditionnel)

Le KWOC aurait alors pu être introduit pour garder le principe d'optimisation spatiale du KWAC tout en ménageant un meilleur confort de lecture, en évitant la coupure abrupte des titres. Le mot-clé est tout simplement repris en tête de chaque ligne (comme le KWAC, pour éviter le centrage de type KWIC), suivi du titre dans son entier. Le mot-clé est donc bien comme sorti de sont contexte pour être répété en tête de ligne. Une deuxième forme de KWOC factorise le mot-clé en le sortant encore davantage des lignes de contexte : si les mots-clés sont en général présents dans plusieurs titres, on peut trouver plus claire et plus concise une présentation où le mot-clé n'est donné qu'une fois, en introduction au groupe de titres concernés. On réinvente alors la présentation d'un index traditionnel.

```
SdN06    Document          Semaine du Document numérique
Albi06   Documents         Colloque Documents numériques et interprétation
Albi06   interprétation    Colloque Documents numériques et interprétation
SdN06    numérique         Semaine du Document numérique
Albi06   numériques        Colloque Documents numériques et interprétation
```

Fig. 3 : Exemple de KWOC (au sens traditionnel), première forme (sans factorisation)

Taylor A. (2000). Wynar's Introduction to Cataloging and Classification. 9th edition. Littleton, colo.: Librairies Unlimited. pp. 408-411),

Luhn a inventé le KWIC vers 1958 (Sékhraoui 1995 cite par exemple :

Luhn H.-P. (1959). "Keyword-in-context index for technical literature (KWIC index)", American Documentation, 11 (4), pp. 288-295. Reprinted in Hays, David G. (ed.), Reading in Automated Language Processing, New York, 1966, pp. 159-167).

Le KWOC et le KWAC auraient été imaginés ultérieurement.

[7] (Salton & McGill 1983) ne consacre que quelques lignes à ce sujet. C'est la première présentation (et longtemps la seule) que j'ai eue des KWIC, KWAC et KWOC, elle a ainsi influencé ma réflexion, mais est peut-être trompeuse. En effet, le KWAC est y explicité comme KeyWord And Context, alors que les références données en note précédente parlent de KeyWord Alongside Context. Le tableau illustratif donné en exemple ne concerne qu'un KWIC et un KWAC, mais ce KWAC est un KWOC au sens des références précédentes. Le KWOC n'est pas décrit ni illustré.

D'une manière générale, le KWAC est beaucoup moins connu que le KWOC, lui-même plus marginal que le KWIC.

[8] La ligne KWIC serait occupée un peu plus qu'à moitié si tous les titres étaient de même longueur et si cette longueur correspondait à la largeur de la page. En pratique, comme ces conditions sont *a priori* assez loin d'être réalisées, la ligne KWIC est plus qu'à moitié vide.

Document(s)
Semaine du Document numérique *(SdN06)*
Colloque Documents numériques et interprétation *(Albi06)*
Interprétation
Colloque Documents numériques et interprétation *(Albi06)*
Numérique(s)
Semaine du Document numérique *(SdN06)*
Colloque Documents numériques et interprétation *(Albi06)*

Fig.4 : Exemple de KWOC (au sens traditionnel), seconde forme (avec factorisation[9])

Enfin, le Double KWIC (proposé par Anthony E. Petrarca et W. Michael Lay au début des années 1970, et dont la désignation est peu connue) est à l'origine de ce perfectionnement décisif du KWIC : le tri des lignes non seulement en fonction du pivot (c'est le premier tri), mais aussi en fonction de son contexte. C'est en effet un deuxième tri qui permet les effets de superposition de contexte si utiles. Actuellement, bien des concordanciers se présentent comme des générateurs de KWIC, et en proposent, sans doute sans le savoir, une forme avancée (issue du principe du Double KWIC) qui s'est imposée par sa pertinence.

6. Originalité et apports de la pratique séculaire des concordances

Après son parcours au sein des concordanciers de toutes origines et de toutes formes, le chercheur venant de l'informatique textuelle peut encore découvrir du nouveau en se penchant sur les pratiques de concordances "d'avant l'informatique". Comme dans bien d'autres cas[10], l'identité de désignation -les *concordances* bibliques des moines *vs* les *concordances* produites par un logiciel d'étude de la Bible- masque en fait des réalités sensiblement différentes.[11]

Pointons particulièrement ici une différence qui touche à la définition que nous avons donnée des concordances : l'alignement vertical (sur une colonne) du mot pivot, avec la superposition des contextes ligne à ligne, qui est une caractéristique majeure de la concordance "KWAC", ne se retrouve pas dans la concordance traditionnelle. Faudrait-il alors plutôt reconnaître dans les concordances manuelle un ancêtre du KWIC[12] ? Et comment expliquer que l'idée de cette technique de mise en forme, si essentielle et si utile pour le KWAC, ait échappé à des générations d'érudits et d'experts de l'analyse des textes ? C'est que le parallélisme, induit visuellement par l'alignement et la superposition des contextes dans le KWAC, n'en est pas moins présent dans les concordances traditionnelles, où il se dessine d'une autre façon. En effet, les contextes d'occurrence d'un mot (pivot) sont généralement organisés par des regroupements sémantiques. Or ces regroupements sont bien souvent corrélés à des similitudes de contextes d'usage, notamment de construction, si bien que la lecture des contextes cités sous une rubrique donnée concentre et rassemble des voisinages récurrents. Les effets de parallélisme éventuels sont favorisés aussi par la brièveté voulue des contextes, en général sur une ligne (comme les KWAC). Plus fondamentalement, le principe herméneutique à la base même des concordances, c'est bien celui de rapprochement des "passages parallèles". La concordance met en relation, dans un corpus à intertextualité dense, des parties de texte qui entrent en connivence par l'usage d'un même mot. L'interprétation se nourrit de cette considération simultanée de contextes ressemblants, et la concordance est un outil venant au renfort de la mémoire, qui déjà, consciemment et inconsciemment, tisse ces liens. D'ailleurs, le second grand type de document outil herméneutique qui fait le pendant à la concordance, c'est la synopse qui, comme son nom l'indique, présente côte à côte des passages textuels dispersés dans le corpus mais proches par

[9] Pour mettre en évidence la factorisation à l'échelle de notre exemple, les mots-clés ont été lemmatisés, de sorte à pouvoir grouper sous une même entrée le singulier et le pluriel d'un même mot. On peut bien sûr construire un KWOC avec factorisation et sans lemmatisation, c'est même le cas dans la lignée la plus directe des KWIC et KWAC, où les techniques peuvent être très simples et se limiter à des réorganisations de chaînes de caractères.

[10] Celui de l'indexation ou des mots-clés par exemple, l'indexation automatisée du texte intégral dans les corpus documentaires numériques produisant des mots-clés fondamentalement différents, dans leur nature et leur fonctionnement, de ceux affectés à un texte lors d'un catalogage par un documentaliste.

[11] Dans cette partie nous résumons une étude qui pourra être développée dans une autre publication.

[12] KWIC : au sens que nous lui avons donné en § 3.

leur sens et leur contenu. La synopse et la concordance sont deux points d'entrée d'une même pratique de rapprochements de contextes, la synopse partant du texte, la concordance du mot[13].

Les concordances "papier" traditionnelles et les sorties d'un concordancier diffèrent davantage que par leur mode de production ; mieux, chacune tire le meilleur parti des spécificités de leur processus de construction. La concordance construite manuellement dose le détail des relevés et la synthèse des informations (désambiguïsation et lemmatisation contrôlée, abstraction des constructions grammaticales, regroupements sémantiques, sélectivité...). Des contextes interprétatifs pertinents sont délimités à tous les niveaux (entre les entrées, à l'intérieur d'une entrée, choix des citations)[14]. Stable et globale (alors que le concordancier produit des vues dynamiques et locales), la concordance éditée implique des choix, dont la pertinence suppose une réflexion humaine (et non un traitement machinal). Œuvre d'auteur, elle communique une intelligence du texte, une lecture qui fait autorité. Le concordancier s'inscrit en complémentarité : le calcul assure un relevé systématique, régulier et exhaustif. Les tris et alignements verticaux suggèrent des regroupements des contextes par des parallélismes quelquefois inattendus et révélateurs. Grâce à la puissance et à la disponibilité du calcul, l'analyse textuelle se construit dynamiquement, à la croisée de multiples concordances, en variant les entrées et les tris : l'automatisation de la procédure ne dispense pas l'utilisateur d'une certaine habileté herméneutique, pour trouver un éventail de points de vue pertinents et éviter la dispersion. Plus fondamentalement, les concordances tirent parti de l'écriture en tant que technique d'analyse et support de réflexion, et les concordanciers ouvrent de nouvelles perspectives apparentées aux caractéristiques des traitements numériques[15]. Cependant, par des voies différentes, concordances manuelles et concordances calculées servent le même principe herméneutique fondamental : la mise en évidence des parallélismes et des contrastes dans les contextes de l'item étudié.

7. Une proposition technique d'amélioration des concordanciers : les zones

Pour renforcer encore les atouts spécifiques du concordanciers (qui tiennent aux effets visuels d'alignement, de superposition et de répétition), nous avons proposé d'introduire la notion de "zones" dans la technique de construction des concordances. Nous résumons ici la présentation développée, commentée et illustrée dans (Pincemin 2006).

La sélection du pivot se détaille comme une séquence de zones successives : autrement dit, le pivot n'est pas un bloc, mais il est structuré, et composé d'une suite d'éléments individuellement identifiables et potentiellement actifs pour la construction de la présentation des résultats. L'intérêt de ce découpage en zones est de pouvoir indiquer, pour chacune d'elles, *(i)* si elle est le lieu d'un empilement (en formant une colonne), *(ii)* si son unité est soulignée par une mise en valeur typographique et laquelle (par exemple couleur, gras), *(iii)* si elle fait l'objet d'un tri et dans ce cas sur quelle dimension descriptive et de quel type (alphabétique, hiérarchique i.e. par fréquence décroissante, canonique i.e. suivant un ordre conventionnel). Les contextes gauche et droit disposent également d'une possibilité de tri. Au final le tri de la concordance se définit en fixant un ordre d'application des tris des zones et contextes concernés, s'il y en a effectivement plusieurs.

Un tel concordancier nouvelle génération reprend bien toutes les possibilités de tri proposées dans les concordanciers actuels. Il permet le tri sur des mots distants, tout en maîtrisant mieux la portée des tris. En particulier, le découpage du pivot en zones permet un repérage plus fin que la position en nombre de mots par rapport au pivot, puisqu'on peut par exemple s'appuyer sur l'étiquetage du corpus ou prendre en compte le contexte.

[13] D'autres formes de documents traditionnels trouvent un écho dans nos KWOC, KWAC, KWIC et KWUT. L'*index* s'apparente clairement au KWOC. Les *tables* recouvrent des réalités diverses, s'approchant généralement d'un KWIC avec un contexte minimal. Le *recueil de citations* est également un genre de KWIC, manuel, sélectif et non verbal (l'entrée qui définit un regroupement est une désignation thématique qu'on ne retrouve pas nécessairement littéralement dans les citations).

[14] Ces délimitations de contextes de tous ordres sont véritablement une difficulté pour un traitement automatique, or la contextualisation joue un rôle majeur pour l'analyse du sens d'un mot ou d'un texte.

[15] Nous évoquons ici le concept de *raison graphique* développé par Jack Goody, et son correspondant, la *raison computationnelle*, proposé et analysé par Bruno Bachimont.

Les zones sont donc bien au service d'un bon KWAC : les effets d'alignement vertical sont démultipliés et mieux caractérisés.

8. Vers une compréhension linguistique de la puissance herméneutique des concordances

Le succès persistant des concordances témoigne de leur efficacité et de leur pertinence pratique. Et cette pertinence heuristique se comprend très bien au regard d'une sémantique textuelle et différentielle (Rastier 2001), selon laquelle le sens d'un mot, et plus généralement la construction d'une unité linguistique et son parcours interprétatif, se déterminent à partir de ses contextes de tous ordres, de leur rapprochement et de leurs contrastes. La concordance se présente en effet comme un instrument privilégié de l'étude de multiples contextes :

(i) le contexte syntagmatique, bien sûr, par le voisinage immédiat de chaque attestation du mot étudié ;

(ii) le contexte paradigmatique, par les liens possibles aux diverses éditions, aux traductions et à la formulation originale, par le choix des entrées et l'indication éventuelle de renvois, également par le voisinage ou le regroupement des entrées ;

(iii) les parallélismes synoptiques, par le rapprochement visuel des autres contextes du même mot, éventuellement accentué par des alignements verticaux (superposition en colonne sur le mot commun) et des tris (rapprochement et superposition des contextes identiques et mise en évidence des points de divergence) ;

(iv) le contexte textuel et intertextuel, par la référence qui localise l'extrait cité dans le corpus, et par la perception quantitative de la répartition des attestations dans le corpus.

Notre parcours d'analyse a voulu mettre en évidence des observations ainsi pertinentes pour la conception de concordances. Insistant sur le danger de généralisations confuses, notre cheminement invite à cultiver les atouts propres à la concordance et conjointement à jouer pleinement des complémentarités avec d'autres techniques voisines. Le "bon KWAC" repose fondamentalement sur les effets visuels d'alignements, parallélismes et correspondances, affinés et démultipliés avec le concept de zones que nous proposons de mettre en œuvre dans les concordanciers. Et la complémentarité affirmée articule les diverses formes de relevés d'attestation (KWOC, KWAC et KWIC redéfinis, et prolongés par le KWUT), comme elle motive l'intérêt pour des concordances d'auteur aux côtés de concordanciers bien conçus et utilisés avec méthode.

BIBLIOGRAPHIE

BEHAR, H. 1997. La méthode d'Hubert de Phalèse, *Lexicometrica*, 0, 8 p.

BOURION E. 2001. *L'aide à l'interprétation des textes électroniques*. Thèse de doctorat, Sciences du langage, Université de Nancy II.

CHOUEKA Y. 1984. Conversationnel ou concordances imprimées : le problème de l'exploitation d'un gros corpus, *Informatique et Sciences Humaines*, 61-62, p. 93-105.

LEBART, L., SALEM, A. 1994. *Statistique textuelle*, Dunod.

PINCEMIN, B., ISSAC, F., CHANOVE, M., MATHIEU-COLAS, M. 2006. Concordanciers : thème et variations. In VIPREY J.-M., ed., *Proc. of JADT 2006 (8es Journées internationales d'analyse statistique des données textuelles)*, p. 773-784.

RASTIER, F. 2001. *Arts et sciences du texte*, Presses Universitaires de France.

SALTON, G., MCGILL, M. J. 1983. *Introduction to Modern Information Retrieval*, McGraw-Hill.

SEKHRAOUI, M. 1995. *Concordances : Histoire, méthodes et pratique*. Thèse de Doctorat, Université de la Sorbonne nouvelle Paris 3 et Ecole normale supérieure de Fontenay Saint-Cloud, 509 p.